DOCUMENT RESUME

ED 354 758 FL 020 963

AUTHOR Buell, James G.
TITLE TOEFL and IELTS as Measures of Academic Reading Ability: An Exploratory Study.
PUB DATE 4 Mar 92
NOTE 25p.; Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (26th, Vancouver, Canada, March 4-7, 1992).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Advanced Students; Content Validity; *English (Second Language); *Language Proficiency; *Language Tests; Reading Ability; Scores; Second Language Learning; *Test Interpretation; *Test Reliability; *Test Validity
IDENTIFIERS ACTFL Descriptive Scale for Reading; International English Language Testing System; Test of English as a Foreign Language

ABSTRACT

This paper discusses research conducted in the spring of 1991 that measured the relationship of reading subtest scores to teacher ratings of students' reading abilities. Sixty-eight advanced-level students in an intensive English program took an institutional version of the Test of English as a Foreign Language (TOEFL) and a specimen reading module of the International English Language Testing System battery. The students' reading abilities were assessed by their teachers, using a scale devised by the American Council on the Teaching of Foreign Languages. Reliability estimates were obtained and correlations were run. The research tested the hypothesis that the results of a reading test for academic purposes, based on current theories of test design and construction, would correlate better with teacher observations than would results of a more traditional test (the TOEFL). Data analysis indicated instead that each of the tests correlated moderately well with teacher observations. Although somewhat different patterns of correlations occurred with graduate versus undergraduate students, and with natural sciences majors versus arts and social sciences majors, most differences were not statistically significant. (KM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED354758

TOEFL and IELTS as measures of academic reading ability: an exploratory study

James G. Buell

Paper presented to the 26th annual International Convention of TESOL
Vancouver, Canada
March 4, 1992

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James Buell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL 020963

BEST COPY AVAILABLE

More and more, it has become accepted in our field that there are particular settings in which certain elements of our craft, language teaching, become more important than other elements to our students. William Grabe, in his overview of twenty-five years of ESL reading research, limits his discussion to what he calls academic ESL, putting aside discussion of reading for U.S. language-minority students and adult basic literacy. Within academic contexts, he notes, "reading is probably the most important skill for second language learners" (Grabe 1991, 375-6).

Grabe lists six elements that make up the complex skill of reading in a first or second language. These are automatic recognition skills, vocabulary and structural knowledge, formal discourse structure knowledge, content/world background knowledge, synthesis and evaluation skills or strategies, and metacognitive awareness and skills monitoring (379).

Different approaches have been taken to the testing and evaluation of this important set of skills. For the vast majority of the international students entering American universities today, their reading ability in English is inferred from the scores they obtain on the Test of English as a Second Language, and particularly from their scores on TOEFL's Section Three, "Vocabulary and Reading Comprehension." This is in keeping with the Educational Testing Service's own assessment of what TOEFL is testing. Grant Henning, for instance, has written that "TOEFL is by design a test of English for academic purposes," and that its corpus of items uses "a majority of academically-oriented stems" for its vocabulary (Henning 1989, 221).

It is also true that international students whose first language is not English have another means of entry into American universities. For those who cannot reach the cutoff scores on TOEFL established by their universities of choice, there is the path of entry via an intensive English program, most typically one affiliated with a particular university. In many instances, successful completion of a program of intensive English results in a recommendation that a student be admitted to the university to begin academic studies. Typically, such a student must take an institutional version of TOEFL between graduation from the IEP and entry to the university, yet the criterion of admission is not the score he or she obtains on that test, so much as his or her teachers' assessment that the student is prepared to begin academic work.

It is evident, then, that language-related factors other than scores on TOEFL are important to universities considering whether to grant a student admission. The existence of this alternative method of assessing student readiness can be considered valid, to the extent that we can

concur with Johnson's (1983) assertion (as cited by Cohen 1990, 101) that "the best assessment of [reading] ability consist[s] of having teachers interact with students reading authentic texts for genuine purposes, and have them see how their students construct meaning."

It is worth noting here that the concepts on which TOEFL is based are rather different from those which determine success in the typical intensive English program today. In part, the basis of the distinction is theoretical. Bernard Spolsky (1976; cited in Weir 1990, 3) is among many who have identified TOEFL as an example of the "psychometric-structuralist" variety of language test. Davidson and Bachman (1990, 26), go so far as to label TOEFL "the prototypical 'psychometric-structuralist' test [in that] it draws from structuralist tradition and its approach to language which, as Spolsky pointed out, is readily compatible with so called 'discrete point' psychometric tests." TOEFL, they write, "is a multi-item assessment of highly discrete language tasks."

A different view of language informs models of second language acquisition and learning that many of today's ESL teachers and IEP programs subscribe to. John Oller, for one, has been a powerful proponent of the view that "the whole is greater than the sum of its parts" in language teaching and testing (Oller 1979, 212). In Oller's view, discrete-point teaching and testing encourages the notion that elements of language are separable from one another.

Oller's suggestion that cloze testing and dictation be used as means of assessing overall language ability led to a fair amount of research in the 1970s and early 1980s, but these types of tests have themselves engendered a certain amount of criticism. British researcher Cyril Weir (1990), for instance, observes that "although the tests might integrate disparate language skills in ways which more closely approximate actual language use, one would argue that their claim to the mantle of communicative validity remains suspect, as only direct tests which simulate relevant authentic communication tasks can claim to mirror actual communicative interaction (p. 5).

Weir's comments help put into perspective the claim advanced by designers of the International English Language Testing System (IELTS) battery that their own test, unveiled in 1989, represents "a readily available method of assessing the English language proficiency of non-native speakers who intend to study or train in the medium of English" (<u>An Introduction to IELTS</u>, p. 1). IELTS is designed as an integrative test of language-related abilities which are believed to be especially important for would-be international students at British and Australian universities. Weir (1990, 7-15) calls tests of the IELTS

variety "communicative tests," in that they mirror real-life tasks that language learners may likely have to perform. Mari Wesche, a developer of the Ontario Test of English as a Second Language, offers the label, "academic performance test" for this type of evaluative instrument (Wesche 1985, 1-12).

Examination of the constructs underlying the IELTS academic modules indicates that many of the academic language skills emphasized in university-affiliated intensive English programs are similar to those which IELTS tests for. The 35-item specimen test Module C, intended for examinees who plan to major in the arts and social sciences, consists of item sets that assess an examinee's ability to scan, skim, make inferences, recognize paraphrase, recognize and exploit logical organization, discern main ideas and supporting details, recognize synonymy, and recognize grammatical patterns and vocabulary in context. In terms of test method facets, the main item types are matching, summary cloze, and direct quotation from the authentic reading texts employed in the tests; certain items also employ multiple-choice responses and the filling in of a table. A total of three reading passages are used, with lengths varying from 675 to 950 words.

Interestingly, many of the skills which IELTS tests are similar to those mentioned in guidelines for reading instruction in the upper levels of the intensive English program at Ohio University where I carried out my research. Among these are:

* to comprehend and relate the main ideas of authentic reading passages
* to make use of markers of cohesion and coherence
* to skim for general meaning, to scan for specific information
* to identify details which support or develop main ideas
* to refer to textual evidence
* to recognize, select and manipulate information from texts
* to infer conceptual meaning
* to recognize definition in a text
* to recognize various rhetorical modes and expository types
* to recognize and exploit text-based paraphrase
* to relate academic text to diagrams
* to recognize inferences
* to exhibit understanding of academic journal articles

(Combined Skills Course Curriculum Guidelines, Ohio Program of Intensive English 1991).

Both IELTS specifications and the guidelines established by the Ohio Program of Intensive English contrast markedly with the item types employed in the Vocabulary and Reading section of the Test of English as a Foreign Language. Fully half of the sixty multiple-choice

items on this subtest (items 1-30) are strict tests of word synonymy; although sentence completion is the method used, each distractor choice could fit meaningfully into the stem sentence. Items 31 to 60, also multiple-choice, assess understanding of several short passages on varied topics. Each passage is between about 120 and 250 words in length, with each of the passages followed by between four and eight multiple-choice questions. Most questions are concerned with main and secondary ideas in the passages, although questions based on inferences and analogies are also included.

It was hypothesized, then, that the academic reading module of the IELTS specimen test might be expected to assess the skills considered important to an intensive English program today more accurately than the TOEFL reading subtest. This hypothesis became the focus of research which I carried out in the Spring quarter of 1991, using students in the upper levels of instruction at the Ohio Program of Intensive English. The null hypothesis, H_0, was that scores attained by the 64 students in this sample on the IELTS academic reading module would not correlate more highly with teacher assessments of these students' reading abilities, than would scores obtained on the reading section an institutional TOEFL test taken by the same students. The alternative hypothesis, of course, was that IELTS scores *would* correlate more highly with teacher observations than would TOEFL scores.

I should clarify one or two points about IELTS before going on. IELTS has been designed as an integrated set of tests, and the reading and writing subsections are especially closely linked. That is, in an actual administration of the test battery, examinees finish the 55-minute reading subtest, and thereafter spend an additional 60 minutes on two writing tasks that call, in part, for use of information from the readings. In my research, I used only the 55-minute reading test. A second point that needs to be made clear is that three different IELTS reading-writing modules exist, for the fields of arts and social sciences, life and medical sciences, and science and technology. The test design calls for students who intend to major in different academic fields, to take different forms of the test. In my Ohio University study, all subjects took the same form of the test, a specimen version of the arts and social sciences reading module.

Naturally, my research design called for some means of quantifying teachers' observations concerning their students' reading abilities. Correlations don't work without numbers. Although I had considered having the teachers simply rank their students in ability, this proved unsatisfactory, because there was no means of comparing rankings from one classroom to another. Moreover, rankings

do not indicate the degree of difference between one student's ability and another's -- the difference between the second-best and third-best readers might be much greater, or much less, than the difference between the sixth-best and seventh-best, for example. Research presented by Robert F. Boldt at last year's Language Testing Research Colloquium presented a way out of this dilemma.

Boldt, a researcher with Educational Testing Service, was interested in correlating institutional TOEFL scores with teacher evaluations of students who had taken the test. With dozens of teachers and hundreds of students involved, he needed an instrument that would permit comparable data to be obtained. He chose to use a set of descriptors created by the American Council on the Teaching of Foreign Languages (Byrnes and Canale 1987). Boldt's research presents findings for TOEFL as a whole, and for each of the subtests, as correlated with sets of ACTFL descriptors for various language skills. In my own research, I elected to follow Boldt's design.

The purpose of Boldt's research was to determine whether the ACTFL descriptors could form an "anchor" for narrative descriptions of ranges of TOEFL scores. This would, in effect, permit TOEFL to be used in a similar fashion to the ETS Test of Written English, the ETS Test of Spoken English, and the IELTS, each of which reports its results in terms of band scores attached to narrative descriptions.

Boldt's study found moderate correlations between the TOEFL scores and ACTFL measures. Interestingly, he attributed the absence of very high correlations to problems with the ACTFL descriptors, rather than to possible problems with TOEFL itself. One potential contribution of the research which I am reporting on today, then, is to indicate whether it is the ACTFL descriptors that are imprecise measures of academic reading ability, or whether instead its descriptors might match better with scores on a different variety of academic reading test, such as IELTS.

The foregoing discussion provides the framework for the study on which I am reporting today. My major research questions were as follows:

1. What level of correlation exists between scores obtained on TOEFL Section 3, a "psychometric-structuralist" test of reading ability, and on an "academic performance" test of reading, the IELTS Academic Module C reading test?
2. What level of correlation exists between scores obtained on the two types of reading tests, and teacher ratings of students' reading abilities according to the ACTFL descriptors?

3. Does either set of test scores correlate significantly more highly than the other with ACTFL ratings?
4. Does either set of test scores bias for or against particular subgroups of candidates, according either to their academic fields or their graduate/undergraduate status?

The first three research questions are based on a notion of "discriminant validity" (Bachman 1990, 263; Boldt 1991, 8). The assumption is that scores which are associated with the same underlying construct should correlate more highly than those which are not. Both TOEFL and IELTS are claimed to test a similar broad construct of "academic language ability." However, there is a testable hypothesis that TOEFL's test methods, grounded in a psychometric-structuralist view of language testing, might interfere with measuring the construct of academic reading ability, as it would be interpreted by classroom teachers in an intensive English program today.

The fourth question involves both fairness, and a nod to the existing body of research. If we are to consider replacing an existing test with a new one, we must ensure that the new test at least does not bias against certain candidates in ways that the old one does not. This is particularly the case with our use of a test like IELTS, where the developers have created different modules for different academic disciplines, and where our research design has required students from all disciplines to take a single version of the test. In addition, Clapham (1991) has found evidence for a "background effect," or "subject effect" operating at statistically significant levels for a corpus of several hundred IELTS examinees who took two or more versions of the specimen reading tests.

Methodology

Subjects

Five intact classes at the highest levels of the Ohio Program of Intensive English were selected for study in the Spring term of 1991. These were the four part-time "core" classes," Combined Skills 1, CS 2a, CS 2b, and CS 3, and the highest-level full-time English class, Advanced 1b. (Students at the full-time levels of instruction in OPIE take 25 hours of ESL classes per week in a ten-week term. Part-time students take a ten-hour-per-week "core" reading-writing class, and may take an optional five hours per week of specialized instruction in reading, grammar or listening-speaking. CS students may also take five to ten hours per week of academic courses in Ohio University, and most elect to do so; only in rare instances do full-time-level students receive permission to take an academic course concurrent with their ESL coursework.)

In all, 64 students from the five classes participated in the study's ACTFL and IELTS phases. (Four additional students who were rated and took the IELTS did not sit for the TOEFL with their classmates.) Their number included 25 prospective or current graduate students, 27 prospective or current undergraduates, and 12 "OPIE Special" students who had not applied for enrollment at Ohio University. Of the 52 subjects enrolled at Ohio University, 32 had declared majors in arts or social sciences, in fields such as economics, business, linguistics, fine arts, music, philosophy, political science, psychology and communications. Sixteen had declared majors in the physical sciences or technology, in such areas as computer science, mathematics, chemistry, engineering and physics. The remaining four were identified as "undergraduate college" students without majors.

Typical of OPIE classes, the sample tilted heavily toward speakers of Asian languages (n=53): Koreans (19), Taiwanese (11), mainland Chinese (11), Japanese (7), Thai (4), and Indonesian (1). Rounding out the group were native speakers of Spanish (4) and Arabic (3), with one student each from Israel, Turkey, the Congo, and Cyprus.

The teachers of the five intensive English reading classes in this study were themselves in effect research subjects, from whom both qualitative and quantitative data were obtained. Aside from the fact that each is an experienced ESL instructor and native English speaker, there are few generalizations to be made. Two were full-time OPIE instructors, two were part-time instructors, and the fifth was an OPIE teaching associate in the first year of a two-year master's degree program in Linguistics/TESL at Ohio University. Four of the teachers instructed the rated students for two hours a day in the nine weeks prior to administration of the IELTS test, while the teaching associate instructed her students for one hour per day. Each reported basing his or her evaluation of the students' reading abilities on a combination of factors, including in-class tests and quizzes, writing assignments involving reading, and class participation. While the instructors all indicated that inferring the construct of reading ability on the basis of these observations was a complex matter, each expressed confidence in the ACTFL ratings he or she gave to the students.

Administration of IELTS

Each of the students in the sample took the IELTS reading subtests for arts and social sciences (Specimen Module C) along with his or her class during regular class time and in the students' usual classroom. Tests were administered during the ninth week of class.

So that each of the examinees could be made familiar with the IELTS format, each had been provided with a separate module of the reading subtest (Specimen Module B -- Life Sciences) one or two days prior to administration of the test. Also provided were an answer key for the practice test, and an information sheet, "What is the IELTS?", which I had prepared. Students were informed that the content of the experimental test would be different from that of the subtest, but that the item formats would be similar. Nonetheless, the subjects' relative lack of familiarity with the design of the IELTS test cannot be ruled out as a factor in the results obtained.

For each administration of the IELTS test, identical procedures were followed. I began each session by informing the students that their class was a voluntary participant in a "test of the test," with potential influence on the future course of testing in OPIE. While this procedure departs from classic experimental design, the classroom teachers and I felt that giving this information was important for pedagogical reasons; several of the instructors, in fact, stated that the difficulty of the test might cause some of their students to lose confidence in their language abilities if it were not introduced in this manner. Although each subject appeared to put forth his or her best effort while taking the test, the divulgence of this information might have influenced some students' performance.

Time limits were enforced strictly according to standard procedure for administration of the IELTS subtest, as understood by this researcher. Each subject was given 55 minutes to complete the test.

Administration of ACTFL

The 10-level ACTFL descriptive scale for reading was given to the instructors several days before I administered IELTS to their classes. The scale they received differed from the published version only in that descriptive labels (Novice, Intermediate, Advanced, Distinguished, and Superior) were removed and replaced by the numbers 1 to 10. Each instructor was also given a class list with the students' names, space for rankings, and additional space for them to offer qualitative descriptions of the bases for their ratings. The teachers were told to complete their ACTFL evaluations as near as possible to the dates on which their classes were tested, and none were informed of the test results until they had turned in the evaluations.

It was not possible for estimates of inter-rater reliability to be obtained for any of the subjects in my research, since nearly every student had just one reading instructor. Fortunately, however, Boldt's (1991) research

included enough subjects with multiple raters for him to estimate a reliability (r = .59) for the ACTFL scale. This estimate was utilized in analyzing the data for this study.

Administration of the TOEFL

The final leg of this study's evaluative triad consisted of an institutional administration of the TOEFL test, which each subject was obliged to take in the week following the end of the ten-week term. As with all institutional tests given by OPIE, this administration was conducted in strict accord with standard test conditions, with OPIE staff serving as proctors. No irregularities were reported by the proctors in any of the classrooms where the test was administered. The tests were scored at Ohio University, and results were made available to me by the OPIE program.

Overall, timing was a crucial issue in the research design employed in this investigation, since what was intended was a concurrent validation of tests. For this reason, the IELTS subtest and ACTFL scale were administered within a few days of one another, and as close to the end of Spring term 1991 as was pedagogically feasible, in the judgment of the classroom teachers and myself. It was unavoidable that there existed a gap of ten or eleven days altogether, and four or five days of instruction, between administration of the IELTS and TOEFL tests. Still, evaluators of the research should take this time lapse into account when considering my analysis of the collected data.

Results of this research and related studies

Descriptive statistics for this data set

For the 64 subjects for whom complete data is available, the descriptive statistics indicate that assumptions of statistical normality have been met, albeit within a somewhat restricted range of scores. The data appear below:

	Mean	S.D.	Median	Min.	Max.
IELTS	13.80	4.94	13	5	24
ACTFL	7	1.05	7	4	10
TOFREAD	49.34	5.32	49	34	58
TOEFLTTL	504.3	34.4	507	413	577

The subjects of this research form a more homogeneous and less proficient group than the overall population of 875,897 examinees who took official administrations of TOEFL from July 1987 through June 1989 (mean 518, s.d. 67) (Educational Testing Service 1990, 20). Differences between

the two groups are statistically significant ($t = -3.19$; d.f. 63; $p = .0022$, 2-tailed). Not surprisingly, then, this group is also significantly different from the ETS population in its reading subscores (ETS mean 51.6; s.d. 7.5) (ETS 1990, 20). ($t = -3.39$; d.f. 63; $p = .0012$, 2-tailed). It is not unexpected that the experimental group would differ from the overall population of TOEFL examinees, since this study's subjects are all members of high-level classes at an intensive English program. However, the group's special characteristics should be kept in mind when attempts are made to generalize on the basis of the study.

This study's group also obtained scores on the IELTS Specimen reading test that are significantly lower than those reached by a group of 525 examinees in research presented by Caroline Clapham in 1991 (mean 16.9; s.d. 6.6) ($t = -5.027$; d.f. 63; $p = .0000$, 2-tailed).

I also need to acknowledge that my experimental group's ACTFL ratings are significantly different from the results reported by Boldt (1991) in his comparisons of TOEFL reading subscores and ACTFL ratings for 369 subjects (mean 6.43; s.d. 1.55) ($t = 6.01$; d.f. 63; $p = .0000$, 2-tailed). The difference is of a direction and magnitude that we might expect, since Boldt's research involved students at all levels of several intensive English programs, while the OPIE study involved only students at higher levels of instruction.

The narrow range of ACTFL scores in my study appears to have had a major effect on the outcome of correlations involving their use. They are quite low, not only in numerical terms but in comparison to other research that has been conducted along similar lines (Boldt 1991). Even so, each of the correlations is at a statistically significant level, and so I have elected to use them as a basis for comparing the test results, in accord with my research design.

All of these provisos are intended as warnings to the audience. Although my data are normally distributed within their ranges, these ranges are more restricted than in most published studies involving similar measures.

TOEFL/IELTS Correlations

In a report which recently became available through ERIC, the outcome of trials involving the IELTS battery in Australia and southeast Asia is reported (Griffin 1990). Among the many findings is the first published evidence correlating IELTS and TOEFL scores for a subjects who have taken both tests. For 15 subjects who had taken a TOEFL test prior to taking the IELTS arts and social science reading subject test, a correlation of .879 is reported.

For 18 subjects taking both TOEFL and the IELTS life sciences reading test, the correlation is .704. For 21 subjects taking both TOEFL and the IELTS science/technology reading test, the correlation is .866 (Griffin 1990). "While many of the sample sizes are small . . . ," he writes, "it is clear that IELTS is measuring language proficiency in the same domain measured by similar test batteries."

My own research indicates a rather lower level of correlation between the IELTS arts and social science subject reading module and TOEFL. Here are the results I obtained:

CORRELATIONS BETWEEN TEST SCORES AND TEACHER RATINGS
Using the Pearson Product-Moment Formula

ACTFL				
IELTS	.516(.396)			
TOFREAD	.499(.383)	.554		
TOEFLTTL	.620(.476)	.511	0.844	
	ACTFL	IELTS	TOFREAD	TOEFLTTL

(parentheses indicate r values before correction for attenuation in criterion measure)

In keeping with accepted statistical procedures (Guilford and Fruchter 1973, 441), we partially corrected for attenuation in correlations involving ACTFL. This was done by dividing the observed correlation by the square root of .59, the reliability estimate for ACTFL reported by Boldt (1991). The original, unattenuated correlation is shown in parentheses.

Guilford and Fruchter (1973) also suggest a method of testing for differences between observed correlations when coefficients of correlation come from two different samples. Use of the formula indicates that Griffin's observed correlation is significantly different from our own. ($z = -2.58$; $p<.05$).

Of course, it is one thing to identify a difference between correlation coefficients, and quite another to account for it. It may be that the somewhat restricted range of TOEFL scores in our sample is responsible, though again, this restriction is inevitable given the population of ESL learners around which the study was designed.

IELTS/ACTFL Correlations

Boldt (1991) reports a correlation of .61 between TOEFL reading subtest scores and ACTFL ratings for the subjects in

his research, using the same method of quantification for ACTFL (equal intervals) that I later employed in my research. His data come from the student populations of seven language schools on the U.S. east coast, with a total of 369 students for the comparison of ACTFL and TOEFL reading scores. The correlation I obtained, .38, is significantly different from his own, as computed by the Guilford and Fruchter formula used above ($z = 1.69$; $p < .05$). Likewise, his r_{ca} value of .79 is significantly different from the value of .499 which I found ($z = 2.14$; $p < .05$).

Correlations of the results obtained in my research

Leaving aside the results of others' research for a moment, we now come to the answer to our major research question. That is, is there a significant difference between the correlation we observed for IELTS and ACTFL, and the correlation we observed for TOEFL Reading and ACTFL? In a word, no.

Our answer to this question comes from another formula, Hotelling's test for differences between correlated coefficients of correlation (Guilford and Fruchter 1973, 167). Whether we deal with raw correlations or with correlations as they have been corrected for attenuation, we find a t value of around 0.17, well below the 1.64 needed to overcome the null hypothesis of no difference between correlation coefficients, given a confidence level of .05. As a result, we have to conclude that for our data, there is no significant difference between the correlation observed between ACTFL and IELTS, and the correlation between ACTFL and TOEFL Reading.

Background and maturation effects: the issue of bias

Our fourth major research question involved the possibility that one or more of our tests might favor certain groups of examinees. The IELTS module, in particular, was designed specifically for students intending to major in the arts or social sciences, yet we used subjects with more diverse educational goals. Might this IELTS module have conferred an undeserved advantage on this set of examinees?

It is also evident that IELTS, designed for use with British universities, is intended for a somewhat different population of learners than is served by an American intensive English program. The countries' different systems of higher education might have influence on the age levels, or prior level of education, of students being admitted for study in each country's universities. For this reason, we also needed to check for whether IELTS might have conferred an advantage on graduate or undergraduate students.

Two-way analyses of variance (ANOVA) were run to check for effects due to choice of major or to graduate/ undergraduate status. Our data pool for this set of tests was rather small: 48 of the 64 subjects had declared a major. This number included 32 in the arts and social sciences, and 16 in technical or scientific fields. There were 23 undergraduates and 25 graduate students in the sample.

Descriptive statistics for this subsample of 48 subjects are not significantly different from those of the overall group of 64 subjects (IELTS mean 13.4, s.d. 4.93; ACTFL mean 7.14, s.d. 0.94; TOEFL Reading mean 49.7, s.d. 4.7; TOEFL total mean 504.4, s.d. 31.46).

In interpreting the ANOVA results, we adopted a flexible approach to the question of significance levels, in line with recommendations made by Guilford and Fruchter (1973, 179). Rather than choose an arbitrary cutoff level for alpha of .01, .03, or .05, we elected to suspend judgment in cases where alpha was between .10 and .01. Where alpha levels within this range occur, then, replication of the experiment is recommended. In part, this recommendation stems from the limited size of our sample.

The first of our two-way ANOVAs showed no significant differences in IELTS scores between groups based on major ($f=1.39$; d.f.=1; $p=.24$), status ($f=1.9$; d.f.=2; $p=.17$), or the interaction of these two groupings ($f=1.57$; d.f.=2; $p=.22$). In no case did alpha levels exceed the range of .1 to .10 we set for potential or demonstrated significance.

A second two-way ANOVA was used to test for significant differences in TOEFL reading subscores. There were no significant differences between groups based on major ($f=.317$; d.f.=1; $p=.58$), or on the interaction of major and status ($f=.003$; d.f.=1; $p=.96$). However, the interim alpha range of .1 to .01 was attained in analyzing for the effect of undergraduate versus graduate status ($f=3.07$; d.f.=1; $p=.09$). The alpha in this instance reflects the difference in TOEFL Reading subscore means of 50.89 for undergraduates versus 49.34 for graduate students. This is one area which further research employing a larger sample might profitably examine.

Curiosity led to examination of potentially significant differences between ACTFL scores based on the same criteria of status and majors. Here, the findings were somewhat unexpected, although at levels which force us to be tentative in our conclusions. There was no significant difference between ACTFL scores according to graduate versus undergraduate status ($f=.03$; d.f.=1; $p=.86$) or the interaction of status and major ($f=.90$; d.f.=1; $p=.35$).

However, our interim alpha level of between .1 and .01 was reached in comparisons of the observed variance between arts and social sciences majors and physical sciences majors ($f=4.49$; $d.f.=1$; $p=0.4$). This finding reflects a difference between a mean level of 7.33 for arts and social science majors and the mean of 6.75 for subjects majoring in sciences. For reasons we cannot adequately explain, our ESL instructors appear to have favored arts and social sciences students in evaluating their reading abilities, despite the fact that neither of our two tests, IELTS and TOEFL reading, offers a basis for such a distinction. Again, it should be emphasized that the level of significance is one we prefer to call "questionable." Therefore, we hesitate to offer firm conclusions, and suggest instead that further research be carried out.

Conclusions

Now, we can emerge from the thicket of numbers and formulas, to venture some conclusions about what we have found.

We set out to test the hypothesis that scores on the reading section of IELTS would correlate significantly more highly than TOEFL Reading subtest scores with teacher evaluations of their students' reading ability, as quantified in ACTFL ratings. Instead, we found that the levels of correlation were not significantly different.

On the other hand, the correlation between TOEFL and IELTS tests was itself low enough to suggest that somewhat different abilities were in fact being measured by each test. This last finding is in keeping with the notion that TOEFL, which we have called a "psychometric-structuralist" test of academic reading, is a rather different animal than the newer IELTS, which was developed around a "communicative" or "academic performance" construct.

In the course of examining whether certain groups within our sample might have been put at a disadvantage by the IELTS, we found no significant differences between scores attained by graduates and undergraduates, and between arts and social science majors and physical sciences majors. This finding does not in itself support the IELTS project team's decision to provide separate forms of their test for different groups of majors, especially in light of the fact that, for our sample, TOEFL reading subtest scores for the two groups were not significantly different. Moreover, the finding that our arts and social sciences students may in fact have been more proficient than the others, as judged by the ACTFL ratings which their teachers assigned them, also mitigates against any conclusion that background knowledge played a significant role in IELTS test performance.

In terms of concurrent validation alone, then, this study provides no evidence that either test is superior to the other, in matching teacher assessments of student performance. In neither case are the observed levels of correlation high enough to warrant totally replacing teacher evaluations with test scores, in deciding whether students in the upper levels of an intensive English program are ready to proceed into full-time university studies.

On the other hand, there may well be reasons other than concurrent validation to recommend that an intensive English program like that at Ohio University consider replacing TOEFL with an alternative test. As I mentioned at the outset of my talk, one such reason may be the role of "washback" on instruction. That is, if intensive English students spend inordinate amounts of time preparing to pass the TOEFL hurdle, rather than concentrating on their studies, there is a detrimental effect on the programs. To the extent that "academic performance" tests mirror the goals of upper levels of an intensive English program, it would be wise to consider replacing TOEFL with tests of the IELTS variety.

Naturally, there are caveats involved in this recommendation. First, I need to make clear that the present research involved only tests of reading, and not listening, writing or speaking. Despite the important role of reading ability in academic success, it would be inadvisable to recommend that only tests of reading be given. The TOEFL and IELTS reading subtests are themselves parts of multi-part batteries which attempt to measure, directly or indirectly, proficiency in several areas of language ability.

Whether IELTS itself is the best available means of assessing the academic language ability of would-be entrants to an American university is also open to question. The test battery was, after all, designed to assess readiness to enter British and Australian universities, and attempts to incorporate it into American intensive English programs might be of questionable validity.

Moreover, it must be admitted that the ACTFL ratings themselves could be offering us a somewhat lopsided view of what ESL instructors believe their students' abilities to be. Henning and D... (1990) have published a useful construct validation study that links ACTFL levels to performance on a special battery of tests designed with ACTFL criteria in mind, yet this variety of post hoc validation may not be a fully adequate substitute for a rating scale that is based on empirical evidence from the start. Unfortunately, such a rating scale does not appear yet to exist, although Bachman (1990) does offer hope that one might be developed along the lines of criteria he lays

out in his Fundamental Considerations in Language Testing. Certainly, our finding that ESL instructors might assign ACTFL ratings that confer an advantage on certain groups of students is itself an indication that this particular scale deserves further research.

Finally, recall my earlier cautions about the narrow range of ACTFL scores observed in this study. While the limits are in keeping with what might be expected for students in the upper levels of pre-academic language study, the range is restricted enough that it appears to have been responsible for depressing the correlations involving ACTFL. If at all possible then, replications of this element of the study should employ a scale permitting more variation at this range of ability.

Nonetheless, the questions being raised here are important to teachers and researchers alike. As instructors, it is natural for us to wish for tests that not only discriminate reliably among students, but that are in keeping with our classroom goals. We also hope for evidence, wherever possible, that the results of our tests are in accord with our own evaluations of our students' abilities, although research like this study suggests that there may never be total agreement between teacher-based and test-based assessments.

Our findings suggest several fruitful areas for further research. First, it should be recognized that IELTS is only one of several "academic performance" tests that have been produced to date. The Ontario Test of ESL (Wesche 1985) is one such. A similar battery has been the subject of study at Carlton University, as I learned at last year's Language Testing Research Colloquium. It might therefore be useful for designers and users of these tests to share their results, and even the tests themselves, with one another and with other interested programs in North America. By proceeding in this manner, each of these tests could be further validated, and additional forms might be created.

Admittedly, this avenue of research demands a commitment in terms of time, money and energy far beyond what is entailed by use of an existing, commercially available test. But if the result is tests that satisfy demands of reliability, validity and instructional appropriateness, the effort will be worthwhile.

Reference List

Bachman, Lyle F. 1990. Fundamental considerations in language testing. New York: Oxford.

Boldt, Robert F. 1991. Second language constructs as indexed by ACTFL ratings and TOEFL scores. Draft paper presented as part of the Thirteenth Language Testing Research Colloquium, Princeton, N.J., March 1991.

British Council, Univ. of Cambridge Local Examinations Syndicate, and International Development Program of Australian Universities and Colleges. 1990. International English Language Testing System: Specimen materials for modules A, B, C, general training, listening and speaking. Cambridge: UCLES.

British Council, UCLES, and IDP of Australian Univs. and Colleges. 1990. International English Language Testing System: Specimen materials handbook for the IELTS. Cambridge: UCLES.

British Council, UCLES, and IDP of Australian Univs. and Colleges. 1990. International English Language Testing System: User handbook. Cambridge: UCLES.

Byrnes, Heidi, and Michael Canale, eds. 1987, 1986. Developing and defining proficiency: Guidelines, implementation and concepts (The ACTFL foreign language education series). Lincolnwood, Ill.: National Textbook Co.

Clapham, Caroline. 1991. The effect of academic discipline one reading test performance. Draft paper present as part of the Thirteenth Language Testing Research Colloquium, Princeton, N.J., March 1991.

Clapham, Caroline. 1990. Is ESP testing justified? Draft paper presented as part of the Twelfth Language Testing Research Colloquium, San Francisco, March 1990.

Cohen, Andrew D. 1991, 1980. Testing language ability in the classroom, 2nd ed. (draft manuscript). Boston: Newbury House.

Dandonoli, Patricia, and Grant Henning. 1990. An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. Foreign Language Annals February 1990: 11-22.

Davidson, Fred, and Lyle F. Bachman. 1990. The Cambridge-TOEFL comparability study: An example of the cross-national comparison of language tests. AILA Review 7:24-45.

Evangelauf, Jean. 1990. College Board to revise entrance exams: Says new version will be more useful. Chronicle of Higher Education, 7 Nov 1990: A1, A34.

Grabe, William. 1990. Current developments in second language reading research. TESOL Quarterly 25 (3): 375-406.

Griffin, Patrick. 1990. Characteristics of the test components of the IELTS battery: Australian trial data. Paper presented at the Regional English Language Centre Seminar on Language Testing and Language Program Evaluation, Singapore, April 1990, ED 333 747.

Guilford, J. P., and Benjamin Fruchter. 1973, 1942. Fundamental statistics in psychology and education, fifth ed. New York: McGraw-Hill.

Hatch, Evelyn, and Anne Lazaraton. 1991. The research manual: Design and statistics for applied linguistics. New York: Newbury House.

Henning, Grant. 1989. Comments on the comparability of TOEFL and Cambridge CPE. Language Testing 6 (2): 217-222.

Johnston, Peter H. 1983. Reading comprehension assessment: A cognitive basis. Newark, Del.: International Reading Association.

Ingram, D. E., and Caroline Clapham. 1988. ELTS revision project: A new international test of English proficiency for overseas students. London: British Council, 34 pp. ED 297 570.

Mystat Ver. 1.1, 2.1 (statistical software for Macintosh and IBM-compatible computers). 1990. Evanston, Ill.: SYSTAT, Inc.

Ohio Program of Intensive English. 1988. Combined skills course curriculum guide. Ohio University, Ohio Program of Intensive English, Athens.

Oller, John W. Jr. 1979. Language tests at school. London: Longman.

Weir, Cyril. 1990. Communicative language testing. New York: Prentice Hall.

Wesche, Marjorie Bingham. 1987. Second language performance testing: The Ontario Test of ESL as an example. Language Testing 4 (June): 28-47.

Wesche, Marjorie Bingham. 1985. Introduction. 1-12 in Language performance testing/L'evaluation de la "performance" en langue seconde, ed. Philip C. Hauptman, Raymond LeBlanc, and Marjorie Bingham Wesche. Ottawa: Univ. of Ottawa Press.

Appendix: TESOL '92 Convention Handout

RATIONALE

Scores from the Test of English as a Foreign Language (TOEFL) are a key element in university placement decisions affecting many of the nearly half-million international students in the U.S. today. ESL instructors and researchers are not universally happy with this state of affairs, in part because theories and constructs of language learning and evaluation have changed greatly in the decades since TOEFL first appeared. Concern about the potential for "negative backwash" is common in intensive ESL programs, where some students elect to spend substantial amounts of time preparing for TOEFL instead of for class.

Ubiquitous as TOEFL appears to American eyes, it is not so dominant elsewhere. A recent addition to a line of tests for English for academic purposes is the International English Language Testing System battery, released in 1989 by the British Council, University of Cambridge Local Examinations Syndicate, and International Development Program of Australian Colleges and Universities. Making use of items and tasks drawn from academic settings, IELTS represents an attempt to create a more diverse and direct test of academic language proficiency than the multiple-choice TOEFL.

Instructors could benefit from knowing the extent to which their judgments of students' abilities can be supported by test results. Teachers reluctant to "teach to" the TOEFL might reasonably wonder if results from a different test, designed in accord with current pedagogical approaches, might correspond more closely to their judgments than would TOEFL scores. To explore these questions, this research investigated relationships between scores which 64 upper-level intensive English students obtained on reading sections of the two tests, and between each of these scores and teacher ratings of student reading ability. Teacher judgments were collected as ratings according to the ACTFL generic descriptions for reading.

RESEARCH QUESTIONS

1. What degree of correlation exists between scores obtained on TOEFL Section 3, a "psychometric-structuralist" test of reading ability, and on an "academic performance" test of reading, the IELTS Academic Module C reading test?

2. What level of correlation exists between scores obtained on the two types of reading tests, and teacher ratings of students' reading abilities according to ACTFL descriptors?

3. Does either set of test scores correlate significantly more highly than the other with ACTFL ratings?

4. Does either set of test scores bias for or against particular subgroups of candidates, according either to their academic fields or their graduate/undergraduate status?

METHODOLOGY

Sixty-four students in the advanced and part-time levels of a university-affiliated intensive English program (Ohio Program of Intensive English) participated. At the end of Spring term 1991, each subject took the institutional TOEFL, was rated for reading ability by his/her teacher according to a descriptive scale (ACTFL), and took a specimen version of the IELTS academic reading module in arts and social sciences.

Descriptive statistics were obtained and Pearson correlations were run. Two-way analyses of variance (ANOVAs) were conducted to check for evidence of differences due to academic majors or graduate/undergraduate status.

RESULTS

	Mean	S.D.	Median	Min.	Max.
IELTS	13.80	4.94	13	5	24
ACTFL	7	1.05	7	4	10
TOFREAD	49.34	5.32	49	34	58
TOEFLTTL	504.3	34.4	507	413	577

CORRELATIONS BETWEEN TEST SCORES AND TEACHER RATINGS
Using the Pearson Product-Moment Formula

ACTFL				
IELTS	.516(.396)			
TOFREAD	.499(.383)	.554		
TOEFLTTL	.620(.476)	.511	0.844	
	ACTFL	IELTS	TOFREAD	TOEFLTTL

(parentheses precede correction for attenuation in criterion measure)

Two-way analyses of variance were run for each set of test results. Our data pool (n=48) consisted of 32 arts/social sciences majors and 16 science/technology majors; there were 23 undergraduates and 25 graduate students. Their performance on each measure was not significantly different from that of the overall sample. Taken as a whole, the ANOVAs provide some evidence that these groups were roughly comparable in ability; therefore, the differences which we did find might indicate that one or more of the groups was affected differentially by one or more of the measures, and so they may be worth further study.

IELTS: no significant differences between groups according to major ($f=1.39$; $d.f.=1$; $p=.24$), graduate/undergraduate status ($f=1.9$; $d.f.=2$; $p=.17$), or the interaction of these two groupings ($f=1.57$; $d.f.=2$; $p=.22$).

TOEFL reading: no significant differences between groups based on major ($f=.317$; $d.f.=1$; $p=.58$), or on interaction of major and status ($f=.003$; $d.f.=1$; $p=.96$). However, the alpha attained in analyzing for the effect of undergraduate versus graduate status ($f=3.07$; $d.f.=1$; $p=.09$) is high enough to warrant further study. The alpha reflects the difference in TOEFL Reading subscore means of 50.89 for undergraduates versus 49.34 for graduate students.

ACTFL: no significant difference according to graduate versus undergraduate status ($f=.03$; $d.f.=1$; $p=.86$) or the interaction of status and major ($f=.90$; $d.f.=1$; $p=.35$). However, an alpha level of between .10 and .01 was reached in comparisons of the observed variance between arts and social sciences majors and physical sciences majors ($f=4.49$; $d.f.=1$; $p=0.4$), reflecting a difference between a mean level of 7.33 for arts and social sciences majors and a mean of 6.75 for subjects majoring in sciences. Our ESL instructors appear to have favored arts and social sciences students in evaluating their reading abilities, despite the fact that neither of our two tests, IELTS and TOEFL reading, offers a basis for such a distinction. Overall proficiency, as indicated by total TOEFL scores at least, provides no basis for the difference, since the arts and social science students' scores on TOEFL overall were not significantly higher than the other students'.

LIMITATIONS OF THE STUDY

There are several weaknesses in the study arising from its research design and subject pool. These are:

1. The limited range of subjects' abilities precludes comparison of results with similar studies using broader sampling. Range limitation is of particular concern with the ACTFL ratings, where 57 of the 64 subjects received ratings of 6 (intermediate-high), 7 (advanced), or 8 (advanced-plus) on the 10-level scale.
2. The ACTFL scale used for this study is not based on prior empirical research; there is no certainty that the descriptors represent truly different and hierarchical levels.
3. We have departed from the IELTS developers' intentions in our research design, by giving a subject-specific test to a broad range of examinees.
4. IELTS, designed for British and Australian universities, may contain content unsuitable for decision-making in American university settings.
5. Subjects' comparative lack of familiarity with IELTS format may have influenced their scores.
6. The two-way ANOVA results can provide only an indirect indication of bias, since each set of scores was analyzed separately; the research design and sample size appeared to preclude use of more inclusive measures.

CONCLUSIONS

The relatively weak correlation between TOEFL and IELTS reading scores indicates that separate, although related, abilities are being tested. The degree of relationship between teacher ratings of student

reading ability and TOEFL reading scores in this study is not significantly different from the degree of relationship between the teacher ratings and the students' IELTS reading scores. However, the apparent intrusion of other factors than reading ability into the ACTFL ratings, as indicated by the degree of correlation between ACTFL and total TOEFL scores, casts doubt on whether the ACTFL scale is an entirely appropriate means of quantifying teacher judgments of reading ability.

Equally at issue is the extent to which teacher observations truly reflect student abilities in reading, no matter what descriptive measures are used; if test scores are only indirect measures of ability, the same can be said of teachers' in-class observations of the interaction between readers and texts. Consequently, intensive English programs should select and rely on tests that are in accord with their instructional goals and purposes, and verify that the results correlate acceptably with teacher judgments; yet they should also recognize that neither test results nor teacher judgments is likely to offer a complete picture of ability, and should therefore take both types of information into account.

Several research limitations were described above. With regard to Point 1, it can be argued that the range limit is a reasonable one, given that this is the range of students found in the levels of an intensive English program where decisions about readiness are generally made. As to Point 2, ACTFL ratings are a readily available means of quantifying teacher judgments, if an imperfect one, about which a body of research is being developed. Points 3, 4, and 5 related to weaknesses that could be expected to have decreased the value of correlations between teacher judgments and IELTS, while not affecting correlations between those judgments and TOEFL. Despite these weaknesses, TOEFL and IELTS scores correlated equally well with the ACTFL ratings. This indicates that intensive English programs and test developers would do well to carry out further research into academic performance tests which overcomes the limitations of this study.